Dimensionality Reduction, Clustering, and Neural Network Applications on Spam and HTRU2 Datasets

Dataset Overview

Spam Dataset:

Classifying spam data has been becoming increasingly important as the use of the Internet has scaled. Spam messages for advertising-related, adult-related, and unwanted marketing messages have been rising annually. Consequently, countermeasures such as anti-spam and built-in spam detection in popular email hosts have been flagging and classifying spam appropriately.

The spam dataset analyzed in this paper was selected from the UCI data repository. The dataset uses 57 attributes over 4601 instances (also referred to as samples) to characterize spam data for a user named George Forman. Furthermore, 54 of the 57 attributes (also referred to as features) are a measure of the frequencies of a particular word or character in an email. To name a few, words such as "George, make, address, all, and money" and characters such as "$, #, !" are all counted and calculated as a percentage of the words and characters over the entire email. From this binary classification dataset, 39.4% of emails are truly spam, while 60.6% are not spam.

HTRU2 Dataset:

The HTRU2 dataset, taken from the UCI data repository, lists pulsar candidates collected by the High Time Resolution Universe Survey. Pulsars are a type of neutron star that produce radio emissions detectable on Earth. Pulsars prove to be a unique and highly discussed phenomena in astrophysics.

The HTRU2 dataset uses 8 attributes which are related to the profile or shape of a neutron star. In total there are 17,989 samples, 9.56% of which are positive examples (classified as a pulsar), and 90.84% are negative examples (samples which are not a pulsar).

Dimensionality Reduction and Clustering Overview and Methodology

Dimensionality reduction is a process used to reduce the dimensions in a dataset while keeping those that are most important. In this report we will use principal component analysis (PCA), independent component analysis (ICA), random forest (RF) classifiers, and random projection (RP) dimensionality reduction processes and assess the results of each. After the dimensionality reduction algorithms are performed, clustering will be done on each data set. K-Means and Gaussian Mixture Model (GMM) clustering algorithms will be evaluated for accuracy, efficiency, and generalizability. The GMM algorithm used in this report implemented a "full" covariance type, where each component has its own general covariance matrix, to allow for more flexibility when clustering.

A grid search across feature dimensions for both datasets will be used when evaluating dimension reduction. The values in the grid search set include (N=2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55). Values greater than the dimensionality of the original dataset were not considered in the analysis of this report. Dimensionality reduction scores were calculated and plotted for each reduction technique in order to choose an optimal feature reduction quantity; these scores will be discussed in the subsequent section of the report. A grid search was also done when clustering. The grid search set includes (K=2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40) clusters. Bayesian information criteria (BIC) scores, sum of squared error (SSE), silhouette scores, adjusted mutual information, and clustering accuracies were recorded for each iteration in the grid search. These scores were then evaluated to find optimal cluster quantity based on accuracy and generalizability.

Dimensionality Reduction and Clustering Analysis and Results

Benchmark

In order to create a benchmark, clustering was performed on unprocessed spam and HTRU2 datasets. The results of the clustering across the grid search can be found in Figures 1 and 2 below.

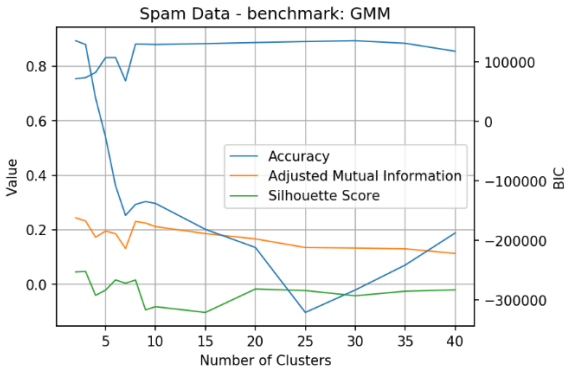


Figure 1. Spam Data GMM Analysis - Benchmark

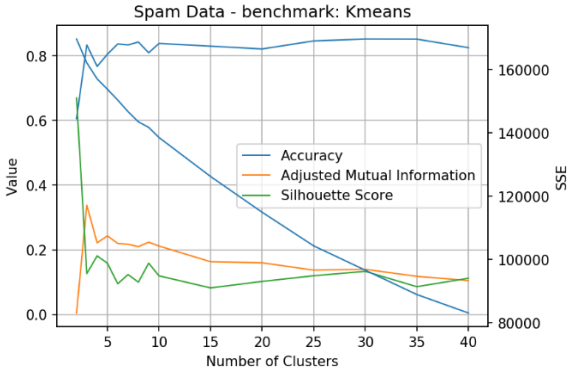


Figure 2. Spam Data K-Means Analysis - Benchmark

Figure 1 above shows that clustering accuracy for the GMM algorithm converged at K=8 clusters. However, in order to find the optimal clustering number based off the results above, a high silhouette score and a low BIC score is desired. A high silhouette score represents a larger discrepancy between clusters, which is desirable because there are typically a set of underlying features which provide separation in the data for classification. BIC score represents the likelihood of the data points belonging to a certain cluster while penalizing model complexity. A low BIC score means data points are more likely to belong to the clusters they are assigned to while maximizing generalizability of the classification. To maximize silhouette score and minimize BIC score without compromising accuracy, a cluster amount of K=25 is assumed to be the optimal number of clusters for data set given this algorithm.

Figure 2 above illustrates that the accuracy converged at K=5 clusters under the K-means algorithm. Like GMM, in order to optimize clustering while preserving accuracy, a low SSE score is desired while maximizing silhouette score. Consequently, it is thought that K=40 clusters is the optimal clustering for this dataset under this algorithm.

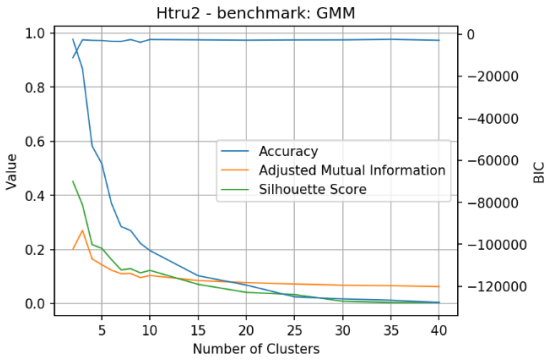


Figure 3. HTRU2 Data GMM Analysis - Benchmark

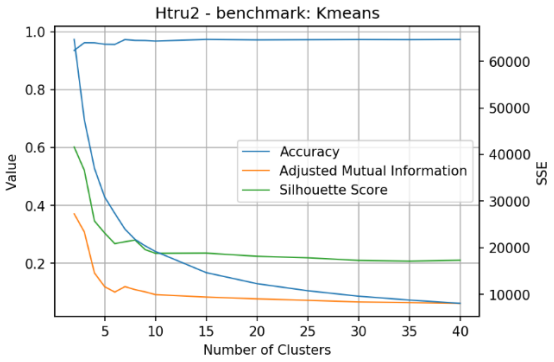


Figure 4. HTRU2 Data K-Means Analysis - Benchmark

Using the same evaluation methods used on the spam data set, the HTRU2 dataset was evaluated to find optimal clustering. Results are illustrated in Figures 3 and 4 above. Silhouette scores and BIC scores converge at K=25 clusters. Furthermore, accuracy converges at K=3 clusters. In order to reduce model complexity while maintaining accuracy and compromising between silhouette and BIC scores, it is thought that K=25 is optimal for the HTRU2 dataset under the GMM algorithm. Similarly, for the K-means algorithm, K=40 is thought be optimal since this maximizes the discrepancy between silhouette score and SSE score while maintaining accuracy.

To visualize the clustering of both datasets, a t-SNE (t-distributed stochastic neighbor embedding) chart was generated; the t-SNE algorithm puts similar data points nearby one another and projects it on a lower dimensional space (2D in this case). The results can be seen in Figures 5 and 6 below.

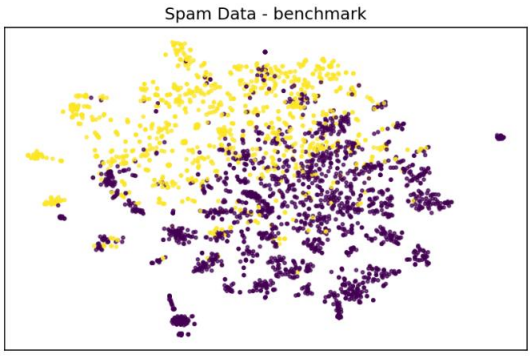


Figure 5. Spam Data t-SNE - Benchmark

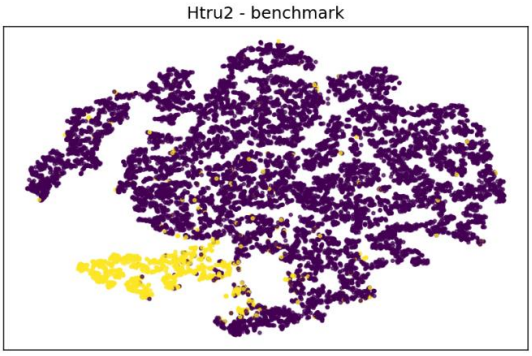


Figure 6. HTRU2 Data t-SNE - Benchmark

The t-SNE chart in Figure 5 illustrates that spam data generally had good separation between classification labels. There are points between the clusters that are intermingled which indicates that there are still datapoints where the classification is unclear. Figure 6 above shows that the HTRU2 clusters are clearly separated, which complements the clustering algorithm's high accuracy.

Independent Components Analysis (ICA) – Dimensionality Reduction and Clustering

ICA was done on both datasets, and the optimal number of components was selected using a measure of kurtosis vs the number of components. A high kurtosis value indicates that the data is more non-gaussian and there is more separation between clusters. Generally, we look for high kurtosis while considering model complexity in order to prevent overfitting on irrelevant features. A "knee" is looked for in a chart (kurtosis vs number of components) to select the optimal number of features. The charts can be seen below in Figures 7 and 8.

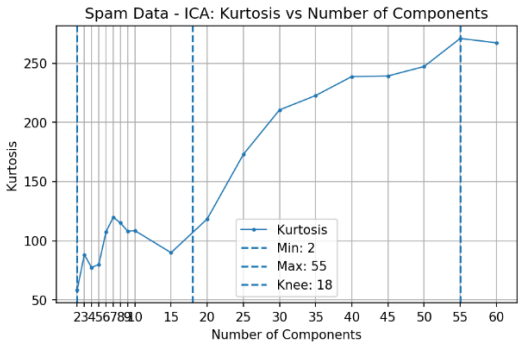


Figure 7. Spam Data ICA Analysis

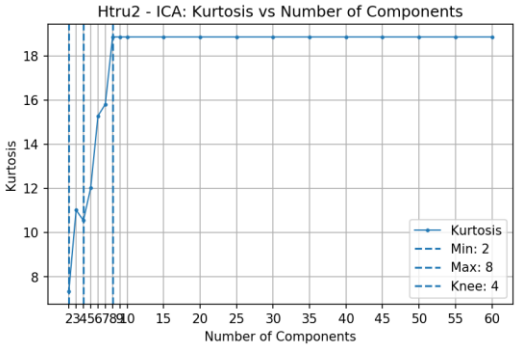


Figure 8. HTRU2 Data ICA Analysis

Figures 7 and 8 show that our search algorithm found knees at K=18 and K=4 for the spam and HTRU2 data sets respectively. However, since K=18 is low in kurtosis relative to K=55, a component value of 55 was selected for the spam data set. Next, clustering was performed on the ICA data to find the optimal number of clusters post-dimensionality reduction. The results can be seen in Figures 9, 10, 11, and 12 below.

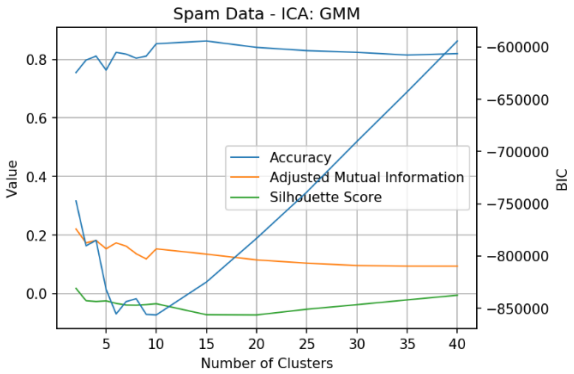


Figure 9. Spam Data GMM Analysis - ICA

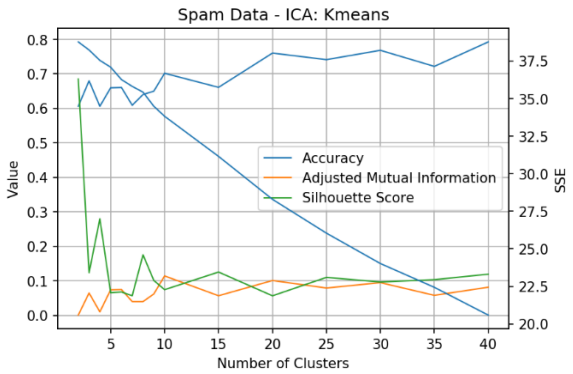


Figure 10. Spam Data K-Means Analysis - ICA

The analysis plots in Figures 9 and 10 above show that the optimal clustering for the Spam Data under GMM and K-means clustering methods is K=10 and K=40 respectively. Clustering selection was done with the same method done on the benchmark outputs, where BIC and SSE scores are minimized while maximizing silhouette scores and preserving accuracy. GMM outperformed the K-means algorithm in terms of accuracy and scored similar to the benchmark clustering. The K-means algorithm generally underperformed from the benchmark. Since the K-means algorithm does not use a covariance matrix or use clustering probabilities for each point in the dataset, it may have allowed for more misclassified data points to influence other data points.

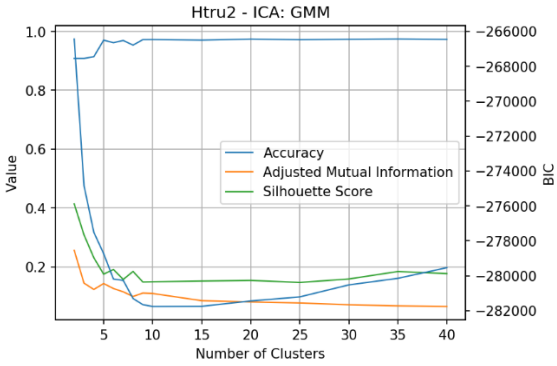


Figure 11. HTRU2 Data GMM Analysis - ICA

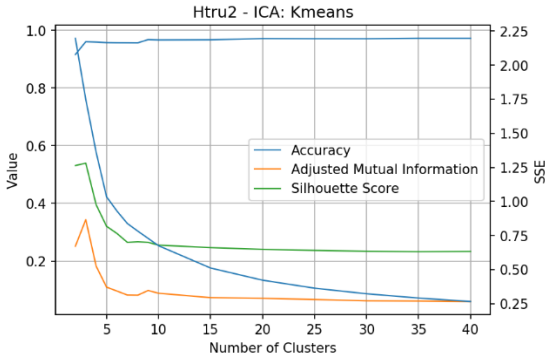


Figure 12. HTRU2 Data K-Means Analysis - ICA

Similarly, for the HTRU2 data set, clusters K=9 and K=40 are thought to be optimal for the GMM and K-means algorithms respectively after analyzing Figures 11 and 12 above. ICA resulted in a high clustering accuracy, much like the benchmark, signifying a successful feature reduction.

T-SNE charts for both datasets are shown below in order to further visualize clustering. The results can be seen in Figures 13 and 14 below.

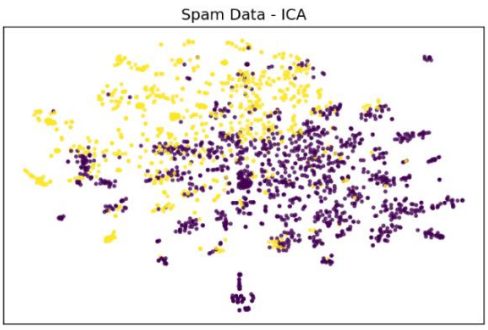


Figure 13. Spam Data t-SNE - ICA

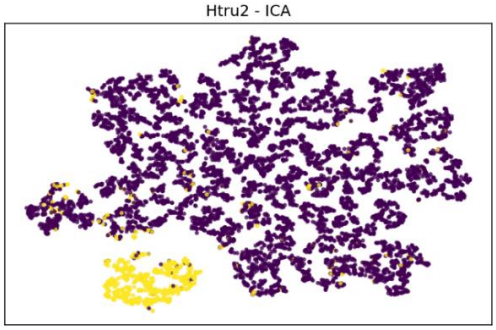


Figure 14. HTRU2 Data t-SNE - ICA

t-SNE charts looks similar to the benchmark results analyzed previously. While accuracy for ICA was generally lower for the spam dataset, there isn't much of a notable difference from the benchmark clustering visual.

Principal Components Analysis (PCA) – Dimensionality Reduction and Clustering

PCA was done to minimize the variance between the datapoints from the principal axes. The charts below plot explained variance, an indirect measure of the number of eigenvalues, against the number of components. Similar to ICA, knees were found to reduce the number of features while preserving important components for clustering.

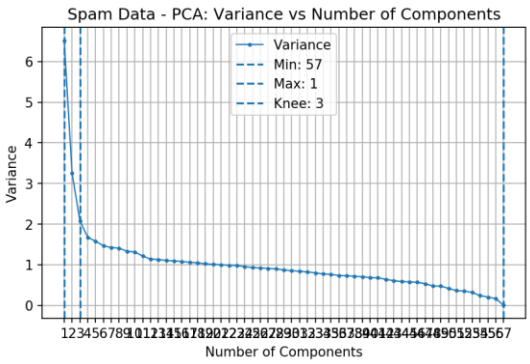


Figure 15. Spam Data PCA Analysis

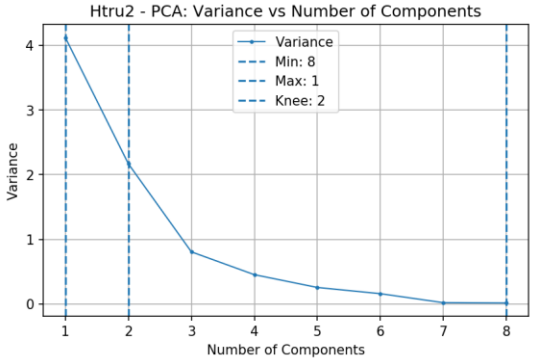
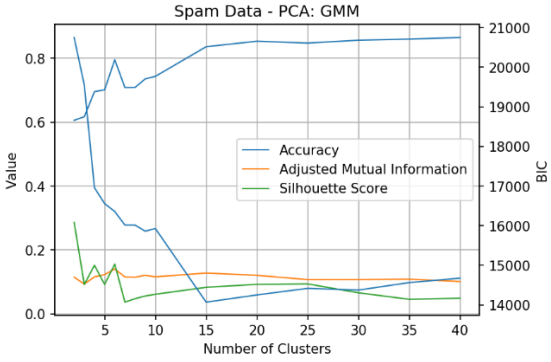


Figure 16. HTRU2 Data PCA Analysis

Figures 13 and 14 show knees were found at N=3 and N=2 for the spam and HTRU2 datasets respectively. However, in the case of HTRU, in order to further minimize variance and guarantee higher accuracy, a knee of N=3 was selected for clustering. The results of the clustering for both datasets can be found in Figures 17, 18, 19, and 20 below.



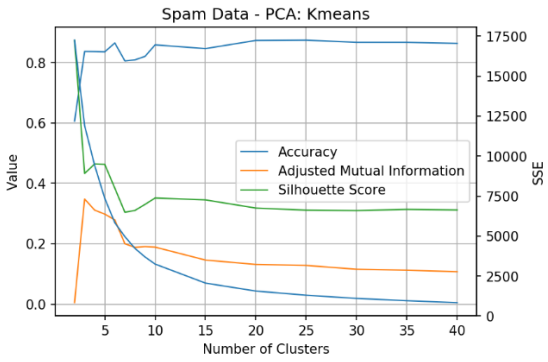
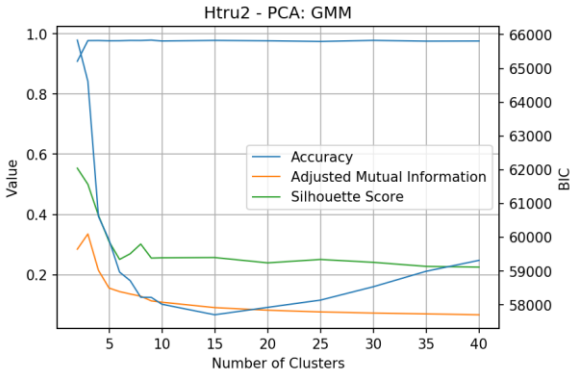


Figure 17. Spam Data GMM Analysis - PCA

Figure 18. Spam Data K-Means Analysis - PCA

Figures 17 and 18 indicate an optimal cluster number of K=15 since silhouette score and BIC/SSE score discrepancies were maximized while maintain accuracy. Clustering was performed with similar accuracy as the benchmark, indicating successful clustering using 3 new dimensions instead of the 57 original dimensions.



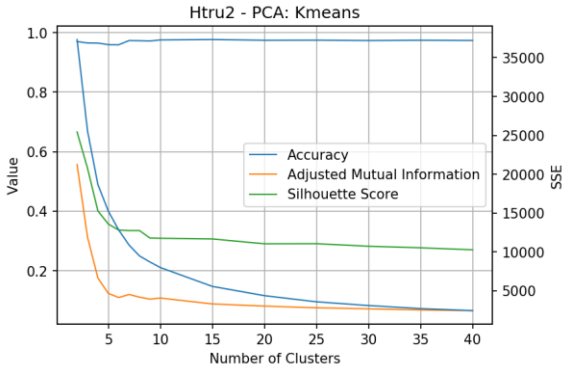
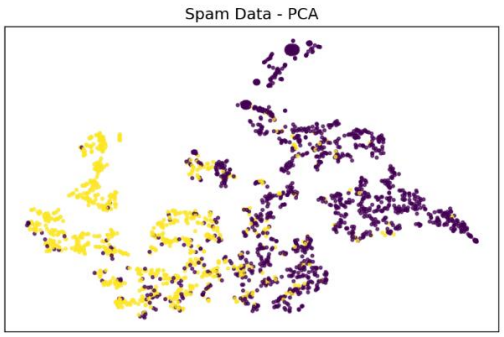


Figure 19. HTRU2 Data GMM Analysis - PCA

Figure 20. HTRU2 Data K-Means Analysis - PCA

Figures 19 and 20 indicate an optimal cluster value of K=15 and K=40 respectively. These clusters maximize the discrepancy between BIC/SSE scores and silhouette scores while maintaining high accuracy.

Both datasets and clustering algorithms had similar performance to the benchmark clustering. This indicates that PCA was successfully able to reduce the number of dimensions in the dataset while providing variance amongst the data points. The results of the clustering can be seen in the t-SNE charts below (Figures 21 and 22)



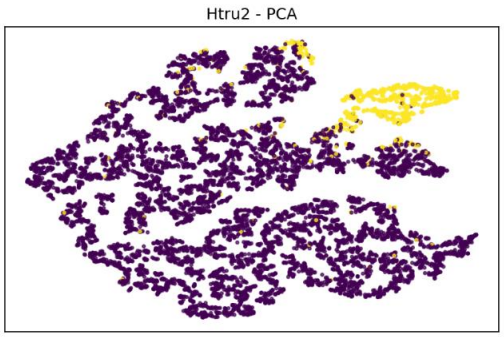


Figure 21. Spam Data t-SNE - PCA

Figure 22. HTRU2 Data t-SNE - PCA

t-SNE charts in Figure 21 and 22 indicate that the clustering and organization of data have changed, but as mentioned above, accuracy remained similar.

Random Forest (RF) – Dimensionality Reduction and Clustering

An ensemble of decision trees was used to create a random forest to classify the datasets. Using usage information from tree leaves and nodes, features in the ensemble that provided the least useful information were the first to be removed in dimensionality reduction. Figures 23 and 24 below plot this feature importance against the number of components to select the optimal number of dimensions to remove while reducing model complexity.

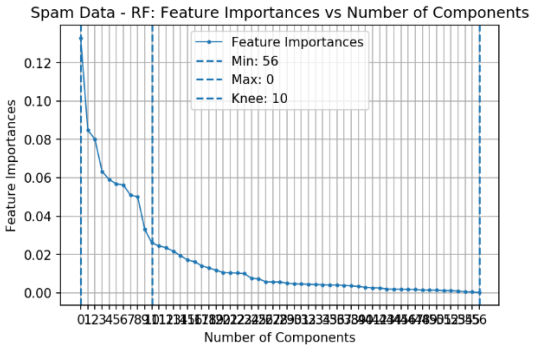


Figure 23. Spam Data RF Analysis

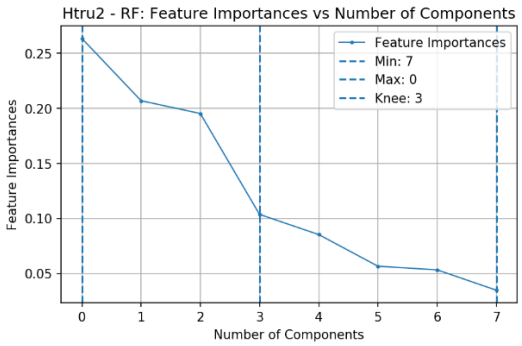


Figure 24. HTRU2 Data RF Analysis

Knees were found at 10 and 3 for the spam and HTRU2 datasets respectively. These values were used for feature reduction and to perform clustering. Clustering results with feature-reduced data can be seen in Figures 25, 26, 27, an 28 below.

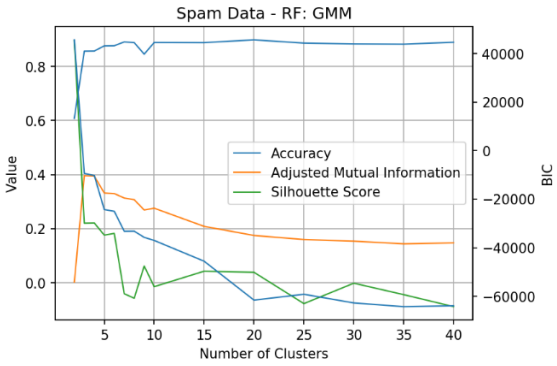


Figure 25. Spam Data GMM Analysis - RF

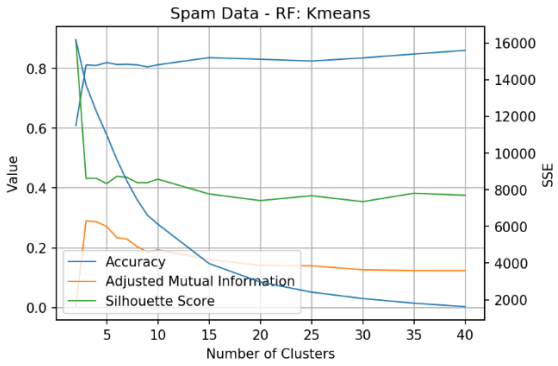


Figure 26. Spam Data K-Means Analysis - RF

Clustering values K=20 and K=40 are assumed to provide the best clustering performance for the spam dataset after random forest dimensionality reduction. These values maximize BIC/SSE score and silhouette score differences while maintaining accuracy.

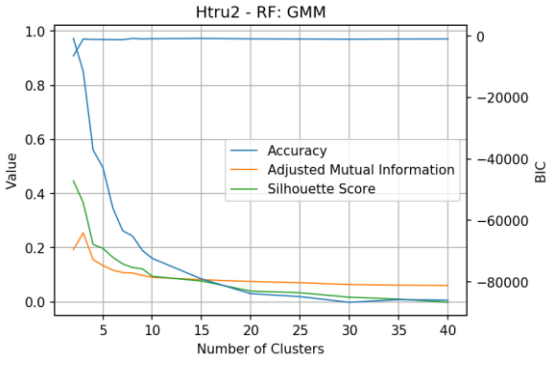


Figure 27. HTRU2 Data GMM Analysis - RF

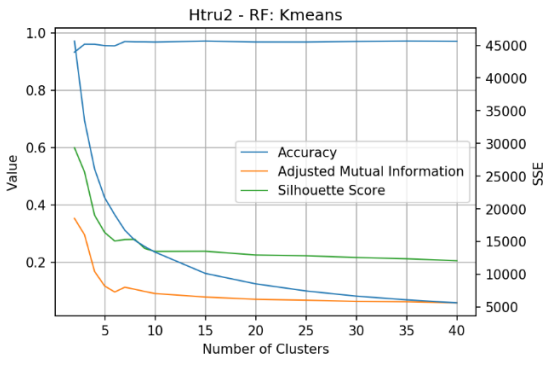


Figure 28. HTRU2 Data K-Means Analysis - RF

For the HTRU2 dataset, K=15 and K=35 are expected to provide the best clustering results while reducing model complexity. These values were thought to be the best compromise between BIC/SSE scores, silhouette scores, and accuracy.

The t-SNE charts below (Figure 29 and 30) show the clustering of data for both datasets.

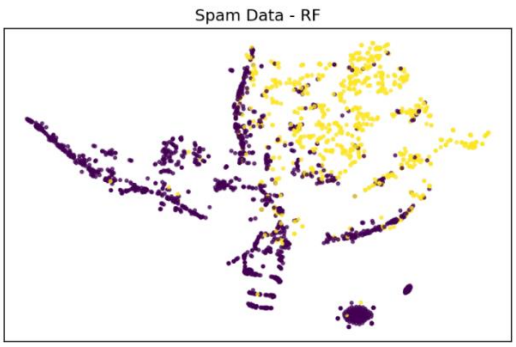


Figure 29. Spam Data t-SNE - RF

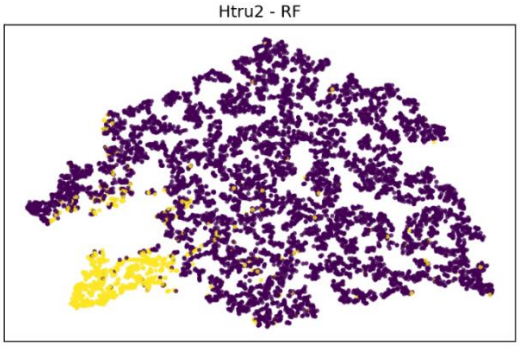


Figure 30. HTRU2 Data t-SNE - RF

The HTRU2 clustering looks similar to benchmark and performed similarly in terms of accuracy. However, the spam data set clustering (shown in Figure 29) looks much different from the benchmark. The points are more spread out relative to one another, indicating that RF feature reduction changed the landscape of the data.

Random Projection (RP) – Dimensionality Reduction and Clustering

Random projections are a dimensionality reduction technique where data is randomly projected in to lower dimensions while preserving distance between any two samples. A pairwise distance correlation coefficient was calculated and plotted against the number of components to evaluate the optimal number of dimensions to remove (shown in Figures 31 and 31). Generally, a higher pairwise distance correlation coefficient is wanted since it represents a higher dependence between dimensions, meaning more of the original dataset information is preserved after dimensionality reduction. However, to limit model complexity and overfitting, a knee is sought after in the plots, similar to ICA and PCA analysis done previously.

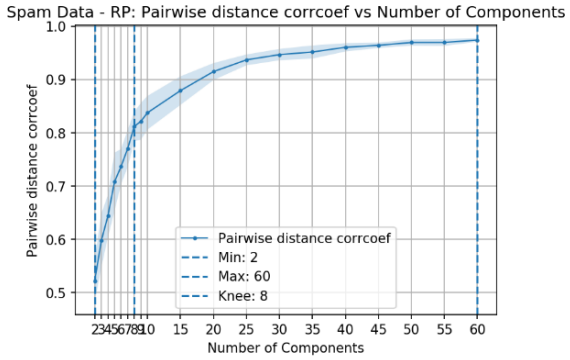


Figure 31. Spam Data RP Analysis

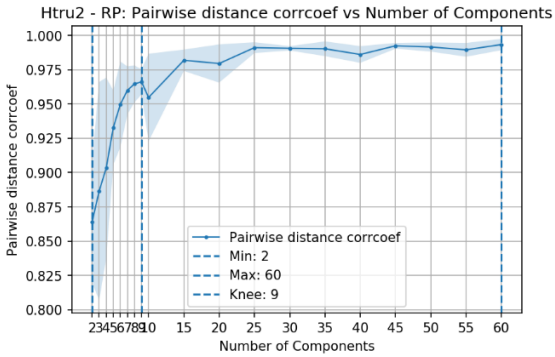


Figure 32. HTRU2 Data RP Analysis

Knees were found at N=8 and N=9 for the spam and HTRU2 datasets respectively. However, since N=9 is in a higher dimensional space than the original HTRU2 dataset where N=8, a component number of N=7 was chosen for clustering.

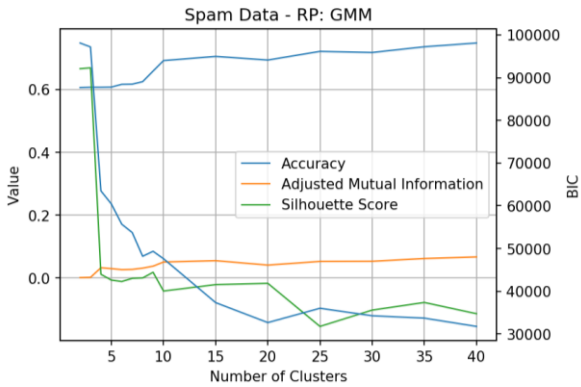


Figure 33. Spam Data GMM Analysis - RP

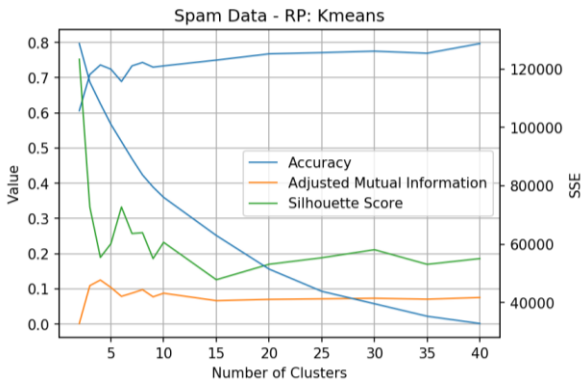


Figure 34. Spam Data K-Means Analysis - RP

After analyzing Figure 33 and 34 above, cluster numbers K=40 and K=20 were chosen from GMM and K-Means analysis respectively. These cluster values maximized silhouette score and BIC/SSE score disparity while maintaining accuracy. However, it should be noted that accuracy was much lower in the GMM model when compared to the benchmark and the K-means algorithm. This is a result of the dimensions that the data was projected on to. The GMM model was not able to accurately classify data because correlation was low since the information contained in the newly projected points were sparse.

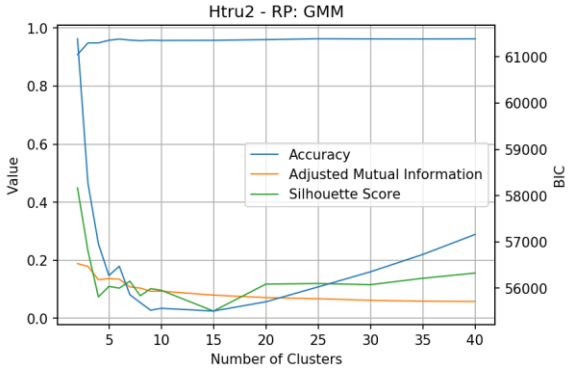


Figure 35. HTRU2 Data GMM Analysis - RP

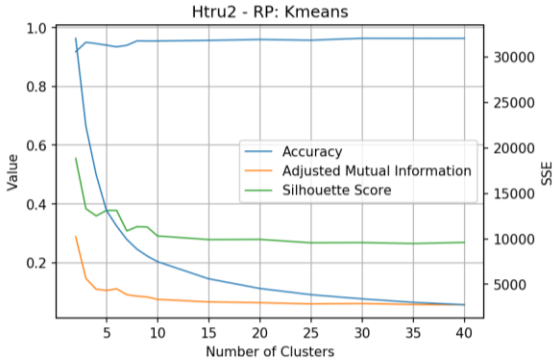


Figure 36. HTRU2 Data K-Means Analysis - RP

Figures 35 and 36 above suggest a cluster value of K=9 for the GMM algorithm and K=40 for the K-Means algorithm. These cluster values try to maximize the discrepancy between BIC/SSE scores and silhouette scores while preserving accuracy.

The t-SNE plots for both datasets can be found in Figures 37 and 38 below.

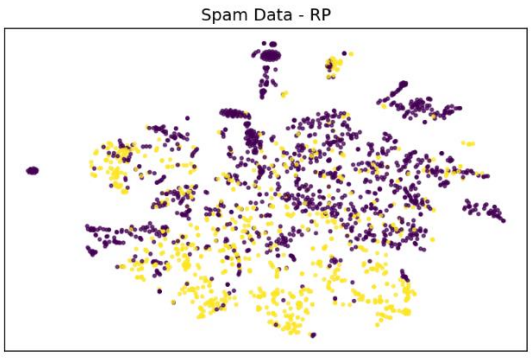


Figure 37. Spam Data t-SNE - RP

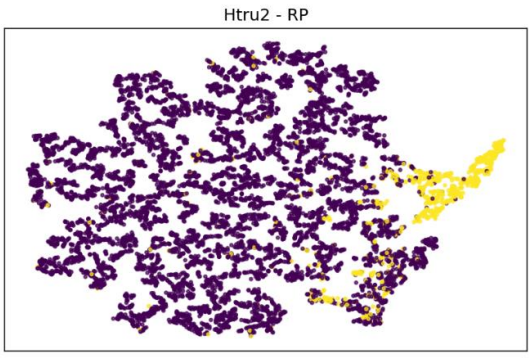


Figure 38. HTRU2 Data t-SNE – RP

Both figures suggest that the clustering has changed from the benchmark cluster. The intermingling of the datapoints in the spam dataset is consistent with the low accuracy since it shows that the algorithm was not able to make a distinct separation between the labeling of the datapoints.

Dimensionality Reduction and Clustering Neural Network Results and Analysis

A neural network analysis was done on the span data set throughout the dimensionality reduction and clustering processes. In the case of clustering, each cluster was treated as a NN input. A neural network grid search was used to tune hyperparameters and maximize accuracy, the results can be seen in Table 1 below. The parameters of the grid search are shown below:

Alpha Values – [1.0, 0.1, 0.01, 0.001, 0.0001, 0.00001]
Layer Sizes – [(50, 50), (50,), (25,), (25, 25), (100, 25, 100)]

Table 1. Neural Network though Dimensionality Reduction and Clustering Processes

	Alpha	Layers	Accuracy	Mean Fit Time (s)
Benchmark	0.001	(100,25,100)	0.9268	0.5332
No DR GMM (K=25)	0.01	(25,)	0.8691	2.4467
No DR K-means (K=40)	0.1	(100,25,100)	0.8897	2.5117
ICA (N=55)	0.00001	(114,114,114)	0.9207	1.7090
ICA+GMM (K=40)	0.001	(50,)	0.8319	7.8833
ICA +K-Means (K=6)	0.1	(25,)	0.5036	0.6117
PCA (N=3)	0.00001	(114,114,114)	0.8700	1.1126
PCA+GMM (K=2)	0.1	(50,50)	0.7017	0.2380
PCA +K-means (K=40)	0.01	(50,50)	0.8794	1.0279
RF (N=8)	0.00001	(114,114,114)	0.8923	1.5038
RF+GMM (K=40)	0.1	(100,25,100)	0.8803	1.7751
RF +K-Means (K=35)	0.0001	(100,25,100)	0.8881	1.9556
RP (N=10)	0.00001	(114,114,114)	0.8522	1.2407
RP+GMM (K=40)	0.1	(100,25,100)	0.7034	2.3118
RP +K-means (K=30)	0.0001	(100,25,100)	0.7775	1.7282

Table 1 suggests accuracies were generally lower when using clusters as inputs for the NN. When comparing dimensionality reduction techniques, ICA yielded the highest accuracy for the knee selected, however, only 2 features dimensions were removed. PCA, RF, and RP methods suggest that up to 54 dimensions can be removed for the NN with a small compromise on accuracy. Timing for running each NN varied for each dimensionality reduction and clustering technique used, but for optimal performance and speed, it was best to just use the original spam dataset for NN classification. Finally, the NN analysis shows that the optimal cluster selection methods used throughout the report may not result in the highest accuracy. The number of clusters selected for the highest accuracy for the NN deviated from our method, indicating that there could be any number of clusters that result in high or similar accuracies when passing the data through a learner for classification.

References

Hopkins, M., Reeber, E., Foreman, G., & Suermondt, J. (1999, January 07). Retrieved February 2, 2019, from https://archive.ics.uci.edu/ml/datasets/spambase

Lyon, R., Dr. (2017, February 14). HTRU2 Data Set. Retrieved March 13, 2019, from https://archive.ics.uci.edu/ml/datasets/HTRU2